REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of thi

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
10/6/2000

3. REPORT TYPE AND DATES COVERED
Final report 8/1/97 – 7/31/00

4. TITLE AND SUBTITLE
Investigating Molecular Recognition Through Large-Scale Analysis of Protein Sequences and Structures

5. FUNDING NUMBERS
#N000149710725

6. AUTHOR(S)
Mark Gerstein

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Dept. of Molecular Biophysics & Biochemistry
Yale University
P.O. Box 208114
New Haven, CT 06520-8114

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217-5000

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This project studied molecular recognition and enzyme function on a genomic scale. A comprehensive database survey showed that most protein functions are carried out by a single fold and most folds carry out only a single function. There are, however, a small number of multi-purpose folds which carry out many functions. Further information is available at http://bioinfo.mbb.yale.edu/genome.

14. SUBJECT TERMS
Genomics, bioinformatics

15. NUMBER OF PAGES
4

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
UL

FINAL REPORT

Contract/Grant Number: **# N000149710725**

Principal Investigator(s): **Mark Gerstein**

PI Institution: **Yale U., Molecular Biophysics & Biochemistry Dept.**

Contract/Grant Title: **Investigating Molecular Recognition Through Large-scale Analysis of Protein Sequences and Structures**

Award Period: 8/1/97-7/31/00

OBJECTIVE: The objective of this project is to study protein sequence-structure relationships through large-scale computational analysis of gene sequences and crystal structure in the databanks. The results of this analysis will be used to help better understand molecular recognition.

APPROACH: A "data-mining" approach was taken where the rapidly increasing amount of data in the publicly accessible databanks was sifted by computational techniques of increasing complexity. The techniques employed will include sequence comparison, structure comparison, packing calculations, molecular simulation, and composition analysis.

ACCOMPLISHMENTS (during entire period of grant):
During the period of the grant I principally worked on the setup of my laboratory. In terms of science, I began to do large-scale database comparison of the protein structures encoded by a number of the recently sequenced genomes, e.g. yeast and E. coli. This work involved extensive recognition of distant homologies to known folds and secondary structure prediction. In particular, I accomplished the following objectives:

* SHARED FOLDS. I have compared the proteins in various major phylogenetic divisions (e.g. plants vs. animals) and a number of the first genomes sequenced in terms of super secondary-structures.

* PREDICTION. Using structure-prediction on the genomes, I found that bacterial genomes have more all-helix super-secondary structures (e.g. more four-helix bundles), eukaryote, more all-strand ones, and archaeon, more mixed ones (e.g. more strand-helix-strand units).

* DATABASE SYSTEM. I have tried to integrate everything I did into a relational database system. I have received equipment grants from Informix and Intel allowing my group to implement a robust and high-throughput system, and we have recently begun designing object-relational schemas to accommodate protein data.

* OPTIMIZE. We have helped optimize high-throughput sample preparation for structural genomics and done retrospective datamining on the results (NAR and NSB papers).

* TREES. We have constructed whole genome trees based on a variety of characteristics (Genome Res. paper)

* EXPRESSION. We have developed a system to analyze whole-genome expression data and relate this to subcellular localization in a Bayesian framework (TIG and JMB paper).

* ANNOTATION-TRANSFER. We have measured the degree to which functional annotation can be transferred as a function of sequence similarity (Wilson et al., JMB).

* LITERATURE. We have put forth a variety of proposals on integrating on-line literature with genome annotation.

CONCLUSIONS: Our initial analyses of genomes have shown that a relatively small number of basic structural parts (i.e. folds and structural superfamilies) are common among all organisms. These parts tend to be metabolic scaffolds, of which the TIM-barrel is an exemplar, that can support multiple functions. They also tend to be highly expressed (in gene-expression studies). Conversely, we have also found unique structural parts in some genomes. With regard to pathogens, these could potentially be useful drug targets.

SIGNIFICANCE: Our studies should help in comparing and understanding microbial genomes, in relating protein function and structure, and in helping with the general progress of structural genomics.

PUBLICATIONS, ABSTRACTS, TECHNICAL REPORTS, PATENTS, AND AWARDS (last 12 months):

D Christendat, A Yee, A Dharamsi, Y Kluger, A Savchenko, J R Cort, V Booth, C D Mackereth, V Saridakis, I Ekiel, G Kozlov, K L Maxwell, N Wu, L P. McIntosh, K Gehring, M A. Kennedy, A R Davidson, E F Pai, **M Gerstein**, A M Edwards & C H Arrowsmith, "Structural Proteomics of an Archeon," *Nature Structural Biology* (in press).

S Balasubramanian, T Schneider, **M Gerstein** & L Regan (2000). "Proteomics of Mycoplasma Genitalium: Identification and Characterization of Unannotated and Atypical Proteins in a Small Model Genome," *Nuc. Acid Res.* **301**:1059-75.

A Drawid, R Jansen & **M Gerstein** (2000). "Gene Expression Levels are Correlated with Protein Subcellular Localization," *Trends in Genetics* (in press).

D Christendat, A Yee, A Dharamsi, Y Kluger, **M Gerstein**, C Arrowsmith, A Edwards (2000). "Structural Proteomics: Prospects for High Throughput Sample Preparation," *Progress in Biophysics and Molecular Biology* (in press).

G Naylor & **M Gerstein** (2000). "Measuring Shifts In Function And Evolutionary Opportunity Using Variability Profiles: A Case Study of the Globins," *Journal of Molecular Evolution* (in press).

M Gerstein & R Jansen (2000). "The current excitement in bioinformatics, analysis of whole-genome expression data: How does it relate to protein structure and function?" *Current Opinion in Structural Biology* (in press).

M Gerstein & F M Richards (2000). "Protein Geometry: Distances, Areas, and Volumes," *International Tables for Crystallography* **22**. (International Union of Crystallography, Chester, UK, in press).

R Das, H Hegyi & **M Gerstein** (2000). "Genome Analyses of Spirochetes: A Study of the Protein Structures, Functions and Metabolic Pathways in Treponema pallidum and Borrelia burgdorferi," *Journal of Molecular Microbiology and Biotechnology* (in press).

A Drawid & **M Gerstein** (2000). "A Bayesian System Integrating Expression Data with Sequence Patterns for Localizing Proteins: Comprehensive Application to the Yeast Genome," *J. Mol. Biol.* **360**: 1077-1093

W Krebs & **M Gerstein** (2000) "The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework," *Nucleic Acids Res* **28**: 1665-75.

M Gerstein, J Lin & H Hegyi (2000). "Protein Folds in the Worm Genome," *Pacific Symposium on Biocomputing* **5**: 30-42.

M Gerstein (2000). "Annotation of the human genome," *Science* **288**: 1590.

J Lin & **M Gerstein** (2000). "Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels," *Genome Res.* **10**: 808-1

C Wilson, J Kreychman, & **M Gerstein** (2000). "Assessing Annotation Transfer for Genomics: Quantifying the relations between protein sequence, structure, and function through traditional and probabilistic scores," *J. Mol. Biol.* **297**: 233-49.

R Das & **M Gerstein** (2000). "The Stability of Thermophilic Proteins: A Study Based on Comprehensive Genome Comparison," *Functional & Integrative Genomics* **1**: 76-88.

R Jansen & **M Gerstein** (2000). "Analysis of the Yeast Transcriptome with Broad Structural and Functional Categories: Characterizing Highly Expressed Proteins," *Nuc. Acids Res.* **28**:1481-1488

A Senes, **M Gerstein** & D Engelman (2000). "Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif correlates with beta-branched residues in a position-dependent fashion," *J. Mol. Biol.* **296**: 921-36.

M Gerstein & C Chothia (1999). "Proteins in Motion," *Science* **285**: 1682-3.

E Brodkin & **M Gerstein** (1999). "'E-biomed' and clinical research," *New Engl. J Med.* **341**:1080-1

P Ross-Macdonald, P S R Coelho, T Roemer, S Agarwal, A Kumar, R Jansen, K Cheung, A Sheehan, D Symoniatis, L Umansky, M Heidtman, F K Nelson, H Iwasaki, K Hager, **M Gerstein**, P Miller, G S Roeder & M Snyder (1999). "Large-scale analysis of the yeast genome by transposon tagging and gene disruption," *Nature* **402**: 413-418.

M Gerstein (1999). "Forging links in an electronic paper chain," *Nature* **398**: 20

H Hegyi & **M Gerstein** (1999). "The Relationship between Protein Structure and Function: a Comprehensive Survey with Application to the Yeast Genome," *J Mol. Biol.* **228**: 147-164.

M Gerstein, (1999) "Building the future of biocomputing," *Nature* **399**: 101

S Teichmann, C Chothia & **M Gerstein** (1999). "Advances in Structural Genomics," *Curr. Opin. Struc. Biol.* **9**: 390-399.

J Tsai, R Taylor, C Chothia & **M Gerstein** (1999). "The Packing Density in Proteins: Standard Radii and Volumes," *J. Mol. Biol.* **290**: 253-266.

M Gerstein (1999). "E-publishing on the Web: Promises, Pitfalls, and Payoffs for Bioinformatics," *Bioinformatics* **15**: 429-431.

M Gerstein, R Jansen, T Johnson, J Tsai & W Krebs (1999). "Studying Macromolecular Motions in a Database Framework: from Structure to Sequence," *Rigidity theory & applications* (ed. M F Thorpe & P M Duxbury, Kluwer Academic/Plenum Publishers), 401-442.